To OO or not to OO?
Revelations from Ontological Modelling of an Archaeological Information System

Paul Cripps[1] and Keith May[2]

[1]English Heritage, Centre for Archaeology, Portsmouth, UK
Paul.Cripps@english-heritage.org.uk
[2]English Heritage, Archaeology Department, London, UK
Keith.May@english-heritage.org.uk

Abstract. This paper describes the process of modelling archaeological data as used within the Centre for Archaeology with the aim of using digital technologies to improve recording techniques and the resultant site documentation. This includes the development of a domain ontological model to describe the data and act as the basis for system design and evolution. The paper will discuss the use of semantic web technologies and place the CfA work on the ontology within the wider sectoral context of archaeological research and other cultural heritage work in England and Europe.
Keywords: Ontology, CIDOC-CRM, Revelation

1. Introduction

The English Heritage, Centre for Archaeology's (CfA) Revelation project identified early in its assessment stage that the CfA was not lacking in information systems. Rather the picture was of a plethora of self-contained and isolated systems that had been designed over the past twenty-five years. Most of the CfA systems were designed to fulfil individual project requirements, but without the overall structure to best facilitate the shared use and maximum interoperability of the data being collected or created.

Further assessment work on the Revelation project produced Data Flow diagrams and Entity Relationship models of the existing CfA systems that helped to give a clearer picture of the baseline state of affairs. But even though some of the mists surrounding the existing systems were dissipated what was left was still a series of rather fragmented data models for each system without a clear method for how best to integrate the data held within each of them.

What seemed to be needed was a better way of expressing where the gaps were, both in and between the existing data models and most significantly showing how those gaps might be filled or bridged by looking at other ways of modelling the data. This raised the possibility of exploring new web-based search technologies such as XML and in particular the opportunities presented by emerging semantic web approaches. At this point attention focused upon the CIDOC CRM (Crofts et al. 2003) and solutions that might be provided by an ontological approach to data modelling.

2. Why Use an Ontology?

An ontology enables the shared understanding of the meanings and relationships between information items. It does this by making explicit the semantic meanings behind the data and terms used in a database or other records, and these semantic meanings can then be made available for computer searching. Mapping data to an ontology should enable both experts and non-experts to search for and re-use data across different domains, by using a common ontology of shared meanings. A primary aim of the CfA ontological modelling project is to prepare the groundwork for producing a new information system that can share data most efficiently across internal data sets but which is most easily searched by other users from outside CfA, either users within EH or from the wider public who may not have familiarity with details of CfA data dictionaries. In addition to considering XML technologies for archiving and dissemination of data in a commonly compatible format there is a desire to make the data most readily searchable online. The semantic interoperability offered by mapping the CfA data to the CRM is intended to give a greater depth for interrogation of data sets beyond the current basic keyword enabled search mechanisms. The CRM ontology should provide better integration and interoperability in systems that use it, thus enabling a much greater semantic depth to searches and the potential for cross-domain searching by researchers both within and beyond the archaeological sector.

"Its seems that the semantics behind a large set of diverse (meta)data structures from a domain with many subdisciplines can be expressed by a coherent formal ontology based on the common conceptualisations of the respective domain experts, whereas the data entry structures themselves often seem to resist merging." (Doerr 2003: 2)

3. Pros and Cons in Using the CIDOC CRM

The CIDOC CRM has evolved from the domain of museums documentation. More recently it has become better known in the wider archaeological sector (as the number of papers in this year's conference would seem to testify). Several factors suggested to the CfA that the CRM would provide a valuable approach for modelling archaeological systems.

- Firstly, and possibly the biggest selling point, is that the modelling approach is based on mapping the knowledge of

the domain experts. There was considerable appeal to archaeologists in an approach that simply asked that existing data be mapped to a more conceptual model for it to be usable.

- The CRM's conceptual framework would be most useful for defining conceptual processes that analyse archaeological data but that could not be easily represented by conventional data modelling techniques (e.g. representing the concepts of phasing and grouping).
- The event based modelling of the CRM suited many core archaeological activities.
- The extensibility of the CRM could allow local extensions of the model while maintaining compatibility.
- The potential to model in ways that could relate archaeological data to other disciplines such as environmental, geological, or agricultural domains.
- Using the CRM for modelling provided the advantages of OO modelling without pre-determining an OO or relational implementation.
- Using an existing ontology such as CRM should provide greater standardisation and interoperability with other data sets.

Although the CRM uses techniques similar to Object-Oriented modelling, mapping data to the CRM does not itself provide a model for implementation of a system. The CfA therefore cannot simply create a completely new system using the CRM. Rather we can use the CRM to model our existing data and use it to understand how we would wish to make this data join together conceptually and semantically in any new system. In this approach we can identify areas of data which currently are not captured or recorded digitally and model them using the CRM. Because of this issue of implementation we also created models using Universal Modelling Language (UML) to represent more detailed data entities and relationships.

In this way we should be able to define a conceptual blue print for how the data in a new system should be structured and inter-relate. Such a use of the CRM has been successfully instigated by Nick Crofts in developing a system for a number of interdisciplinary cultural and scientific resources in Geneva (Crofts 1999). One practical issue with using the CRM that arose early in the project was how to find a way of producing verifiable models for the domain-experts who may be totally unfamiliar with ontologies or the CRM. In the end a number of diagrammatic and text based models were produced. The CfA project also employed a consultant with an archaeology and systems design background to help in over-coming some of the communication issues.

A further issue was that the CRM does not currently come with a simple 'User Guide' so it's application has required some methodological development work which is the subject of much of the rest of this paper.

4. Developing a Methodology for the Ontological Modelling Project

The CRM specification itself does not recommend any particular methodology. After consultation, the approach that was adopted by the CFA was derived from general ontology building methods (Denny 2002) and can be summarized in five main stages:

4.1 Acquire Domain Knowledge

We began by defining the domain limits to be the archaeological work of the CfA. This crucially meant we were not trying to map all archaeological systems to the CRM but rather focusing specifically on work carried out by the CfA. Acquiring domain knowledge principally involved collecting all available systems and procedural documentation and collating what was relevant. There were some initial decisions about how best to deal with areas such as project management and admin which are business processes that other types of data mappings may cover more appropriately.

In the process of talking to the users we also decided to also model the existing data using UML diagrams to help in explaining to people how their specific data can be represented and how it relates to other data entities.

4.2 Organize the Ontological Model

This requires two basic operations

1 identifying the global concepts (classes) that best match the data being created.
2 identifying the properties (the roles and relationships between the classes).

The CRM itself does not contain specific methods for how to formally represent the classes or properties, although the models that are given as examples in the CRM were drawn up using the TELOS data model.

4.3 Flesh out the Ontological Model

It was clear from the start that we would need to make graphical representations to explain the modelling to others within the project team and to the CfA staff whose data we were trying to depict. For diagrammatic representations of the CRM and UML models we drew up draft diagrams using basic Windows based graphical and spreadsheet software. In addition we produced text based descriptive documents giving a more detailed description of each class and property and showing their relationships as depicted in the CRM diagrams. Attempts were made to reach a general level of granularity across the model so that each of the main information areas in CfA could see their activities defined. Some areas of the model, in particular the context recording system, generated more detailed degrees of modelling.

4.4 Check the Work

Considerable day-to-day revisions and re-workings of the models took place based on a number of group discussions with domain experts; workshops and feedback from CRM consultants and by simply checking and re-checking with the CfA data users themselves.

4.5 Commit the Ontological Model

A final version of the model will be verified initially by CRM experts and then disseminated wider to the CRM and archaeological communities. Further plans for publication and dissemination will be agreed as part of a dissemination review at the end of the ontological modelling project. It is hoped that, although it is primarily a CfA based model, the core of the model for the archaeological recording system may find broader usage where appropriate in the wider archaeological community.

5. Defining a Conceptual Framework

It was clear from the assessment stage of the Revelation project that a conceptual model of the data and processes used within the CfA was required. It was recognised that such a conceptual model could be used to harmonise the various data repositories and rationalise data handling with respect to the archaeological process as conducted by the CfA.
In order to build such a model, it was necessary to obtain a clear picture of the existing situation. The Revelation project, had already carried out a number of data modelling exercises and this information formed the basis for subsequent work. Data Flow modelling of the various teams had shown a complex web of processes, many of which were composite processes; the use and exchange of data between teams was convoluted, with excessive double-handling. The detailed Entity-Relationship modelling of the Context Recording System had resulted in a clear picture of the actual data held in an archaeological recording system, and highlighted areas of overlap and redundancy. It became apparent that such approaches to modelling, using formal techniques, are best applied to scenarios where the data holdings and processes are already clearly understood, especially scenarios where there are documented systems in place. By applying such techniques across the various teams that fall within the CfA domain, the results obtained reflected the understandings of each team and did not help to work towards a shared understanding of the domain as a whole. What was needed was some kind of framework that could be used to describe the domain as a whole, providing commonality across the domain, while at the same time allowing the requisite level of granularity to represent the data and processes of each team effectively.
Rather than try to build such a conceptual framework from scratch, it made sense to adopt the CRM ontology produced by CIDOC for the purposes of cultural heritage documentation.

6. An Object-Oriented Approach

As such, an object-oriented approach to modelling was adopted, using UML to visualise data holdings and processes. As UML is not associated with any particular modelling methodology, it was not a restrictive way of working, rather UML could be used to illustrate scenarios exactly as they were described and observed by the domain expert. The object oriented approach was helpful in a number of ways. Firstly, the model becomes event driven rather than static. Secondly, the use of object classes, class inheritance and the idea of the IsA relationship and specialisation proved particularly effective for describing concepts to archaeologists who are eminently familiar with such classificatory schemes. Finally, the CIDOC CRM uses an object oriented approach, so by gathering the fine detail using an object oriented approach, the concepts embodied in the ontology can be directly inherited by the classes created to represent CfA data and processes.

7. The Modelling Process in Detail

Working with an object oriented approach and an existing domain ontology had a number of benefits. Firstly, the approach focussed our minds on identifying patterns within the model, where the same objects appear time and again. Secondly, being an event driven model, it became possible to identify data items which were not the products of events and thus identify any missing events, or gaps, in the model. Thirdly, the use of stereotyping allowed the classes within the model to inherit properties directly from the classes within the ontology without the need for a further mapping to be produced. Finally, the use of meta-entities, groups of object classes which can be treated as a single class, allowed us to define generic patterns of behaviour and implement them as required throughout the model. Pattern identification proved particularly successful in identifying generic activities, some of which were then used as meta-entities such as the Actor pattern relating to activities. It was noted that for most activities undertaken by archaeologists, there is a requirement to record who was involved (e.g. the ‘Excavated By’ field on a context sheet used to record the name of the excavator of a context). Furthermore, where people are involved in activities, their participation is often in a specified role (e.g. a project supervisor, responsible for checking completed context sheets). This pattern can be used to create an Actor meta-entity which can then be used throughout the model.
Looking for gaps across the model enabled us to make a number of data items currently in use more explicit in terms of what they represent. Given an event based model and the requirement for all objects to be the result of events, where objects exist with no obvious chain of events, it is possible to identify a gap in the model. A good example of this is the nature of spot dates. Spot dates are assigned to contexts and are stored as attributes of the context. It is this assignation process that is the key to understanding the spot date; the spot date is not simply a value assigned to a context, rather it represents the culmination of an interpretive act, based on evidence of one sort or another and expert assumption. It is vital to fill this particular gap and to capture the interpretive assignment event in order to attach important information regarding the resultant piece of data, i.e. the spot date.
By using an object oriented approach, it was possible to use the classes described in the object-oriented ontology as abstract classes which could be implemented by classes within the CfA model. In this way the CfA Excavation class, representing the activity of excavating contexts according to CfA procedures, can be seen to implement a number of CRM classes, most obviously E7: Activity, but also E12: Production

Event, as documentation results from the activity, and E6: Destruction Event, as the activity destroys the physical remains. In this way, using strict class inheritance, CfA classes inherit from CRM classes, the CfA classes using CRM classes as stereotypes.

8. Revelations

The use of an object oriented, event-based model gives us the ability to structure our information about the past in such a way as to better reflect our understanding of the world. Our world is made up of objects and events and we are used to working with them everyday in a variety of ways. If we extend this reasoning to the past, we can see that past worlds were also made up of objects and events. Furthermore we can say that it is events in the past which result in remains in the present and it is activities, a specialisation of events, in the present which engage with and investigate the remains of the past. As such, we have two groups of events, one in the past about which we wish to infer, and one in the present which we use to infer. These two groups of events are related by the place in which they occur and any physical remains found at that place.

Events in the present are simply those identified as being undertaken as part the workflow at the CfA. These include excavation, various forms of survey, measuring, condition assessments and classification as a few examples. These are all represented in the model as explicit events, using CRM classes as stereotypes.[1]

Events in the past are the key to the archaeological process. It is this set of events which result in the archaeological record and comprise context formation and depositional events, various geochemical, geological, environmental and biological processes, object production and loss and various construction, modification, use, disuse and destruction events relating to features, sites and structures. It is here that the strength of the model becomes apparent. According to the CRM, events occur at places and there are a number of spatial operators for reasoning about spatial relationships. Furthermore, events have a temporal aspect to their nature and the CRM implements Allen's Temporal Operators for reasoning about temporal relationships. Given the explicit nature of the events in the past within the model, it is possible to use these spatial and temporal operators to build the sequence of events for the site. This is not a change in the way we as archaeologists understand archaeological sites, merely a mechanism for creating documentation which represents our understanding in an explicit manner. While it is appreciated by archaeologists that what we are trying to do is understand the sequence of events which led to the archaeological remains as we find them, our documentation currently records a static view from the present and information external to the system is required in order to make sense of this documentation.

9. Future Directions

The next step will be to take the output from this project as the basis for subsequent systems design in a move towards implementation. Given the nature of the model, there will be a number of issues to resolve moving towards implementation relating to the physical embodiment of any proposed system. Currently, the model is platform independent and could be implemented in a number of ways with respect to hardware and software infrastructure. Some aspects of the model, for example polymorphism or multiple inheritance are not supported by many database engines and decisions will need to be made on how to implement such traits without compromising the conceptual basis afforded by the model. Such issues are considered to be of a technical nature and will require a technological solution which is independent of the model.

The models derived so far will be of use in planning the development of CfA systems. The intention is that once the models have been agreed within CfA and peer reviewed within the wider CRM community they can be made available for other archaeologists who might wish to map similar archaeological resources to the CRM.

10. Conclusions

The process of modelling has been informative. Not only are we now able to describe our data holdings and processes in a way that better reflects the nature of the data, this is also possible in terms of an internationally agreed standard which can be used to relate CfA holdings with those of any other CRM compliant data source. Practically speaking, even if the over-arching programme of systems development associated with the Revelation project is halted for any reason, we still have a holistic model for the domain that can be used to ensure that the sort of piecemeal development which occurs today is carried out within a framework to ensure best practice in terms of data storage, access, manipulation and interoperability.

In terms of the approach, we have found that using an object oriented approach is ideally suited to archaeological data, which can be seen to be primarily event driven by nature and can easily be described in terms of objects. Archaeological objects, both conceptual and physical, can easily be described in terms of OO class hierarchies and this hierarchical approach can be seen to share much in common with the sorts of classificatory schemes found throughout the archaeological discipline.

It should be noted that the construction of such an explicit model of data holdings and processes can involve a significant investment in terms of time. It is vital to capture the knowledge of the domain experts and this can only be done by talking through the scenarios in which they operate in detail in order to build a complete picture. Such interviews and discussions are time-consuming but they provide information essential for compiling the model, particularly information regarding data which often only exists in informal channels surrounding the system but not integral to it.

Notes

[1] One point to note here is that the description of these events as being events which occur in the present is simply a method of distinguishing them from archaeological events, i.e. those which led to the formation of the archaeological record and about which we wish to infer; the present here should be taken to mean our modern era in which archaeologists operate rather than the literal present.

Acknowledgements

The views expressed here are principally our own, but they are based on the work of the project team. We would therefore like to thank Dave Fellows, Anne Greenhalgh and David Robinson for their hard work and advice.

References

Crofts, N., 1999. Implementing the CIDOC CRM with a relational database. *MCN spectra* spring 1999.

Crofts, N., Doerr M., Gill T., Stiff M. and Stead S. (eds), 2003. Definition of the CIDOC Conceptual Reference Model and Crossreference manual. Version 3.4.9. Official release of the CIDOC CRM. http://cidoc.ics.forth.gr/official_release_cidoc.html

Denny, M., 2002. Ontology Building: A Survey of Editing Tools. http://www.xml.com/pub/a/2002/11/06/ontologies.html

Doerr, M., 2003. The CIDOC CRM – an Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine*, Volume 24, Number 3. (2003).